

Computational Reinforcement Learning using Rewards from Human Feedback



Syed Ali RAZA

Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2018

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Syed Ali RAZA

Date: 18/July/2018

Abstract

A promising method of learning from human feedback is reward shaping, where a robot is trained via human-delivered instantaneous rewards. The existing approach, which requires numerous reward signals about the quality of agent's actions from the human trainer, is based on a number of assumptions about human capabilities. For example, it assumes that humans can provide a precisely correct feedback to an agent's action, or that they would always prefer to train an agent by means of reward signals, or that they can assess an agent's actions for any length of training.

In this thesis, we have relaxed these assumptions and have addressed two important issues which are not handled by the existing approach. First, how to compute a potential function using human feedback which can indicate the correctness of an action in terms of increasing or decreasing potential. Second, how to design training methods which cater to human preferences. Furthermore, we have identified that there are two important preferences of a human trainer in the application of reward shaping: (a) a preference to transfer knowledge by providing demonstrations and (b) a preference for short training durations. To address these issues, we have introduced three new methods of computing rewards from human-feedback.

The first method, named rewards from state preference, takes human feedback as preferences of states in terms of distance to the goal state. It removes the assumption of highly accurate evaluative feedback from the user. It computes a high-quality potential function for potential-based reward shaping from only a few human feedbacks. Using feedbacks as state preferences, a ranking model is learned which computes a complete ranking of states. These state rankings define a potential function for

potential-based reward shaping. This method learns a policy much faster than a reinforcement learner which is trained without human feedbacks.

The second method, named rewards from action labels, replaces the traditional evaluative-style feedback approach with a demonstration-style feedback approach. The method caters to the human preference of providing a demonstration. It takes human-feedback as an action label for the current state, which is similar to providing demonstrations. The agent acts using its own policy. A reward function is computed by comparing agent's action with the action label. We found that this method can be favorable to a naïve user as compared to the traditional evaluative-style feedback method.

Finally, the third method, named rewards from part-time trainers, is designed to reduce the load of a single dedicated trainer by curtailing the length of a training session. A policy is taught by a number of trainers. Each trainer provides reward signals for a small number of steps. Experiments, using online crowd, showed that the random part-time trainers can collectively train a good policy. In a survey, conducted for this method, people overwhelmingly voted in favor of the idea of training for a short duration.

Overall, this thesis contributes towards further enhancing the application scope of reward shaping. It develops three new efficient techniques of conducting reward shaping using human feedbacks of different types.

Acknowledgements

First of all, I am thankful to Allah for His blessings, for the knowledge He imparted to me, and for everything.

I am extremely thankful to Prof. Mary-Anne Williams for the supervision and support she has provided throughout my candidature. I found her a courageous and lively supervisor. I am thankful to her for exposing to me a broader view of research, academia, and industry. I will be benefiting from her views in the rest of my life.

I am thankful to Prof. Sajjad Haider for his support which enabled me to begin this journey. The technical and non-technical knowledge I gained from him helped me to solve numerous problems. Specially, my technical writing style is greatly influenced by his writing style.

I am thankful to my lab members, all of them contributed something directly or indirectly to my research. Specially, I would like to thank Ben, Xun, Rony, Shaukat, Pramod, Anshar, Wei, Nima, Mahya, Sylvan, Jon, Meg, and Sam. A special thanks to Richard for proofreading this thesis and providing helpful suggestions.

I am thankful to the Australian nation and government for providing me this opportunity. It provided an exposure to world class technologies and facilities. The love for science and research here was a key motivating factor to me. Besides, I enjoyed spending time with the people, I met here, and building relationships.

I am thankful to my parents for enormous invisible support they provided. Instead of advising me to become a money making machine, they have always provided unconditional support to fulfill my desire of higher studies. My younger brother, Hasan, who helped me in supporting family in this duration. I am also thankful to my wife, Munazzah. Because of her, my productivity increased many times.

I am thankful to my friends, specially, those who instigated in me an urge to do a PhD. I am thankful to my friends in Sydney with whom I made countless fun-filled trips.

Finally, I am thankful to everyone who contributed something to this thesis.

Contents

Certificate of Original Authorship	I
Abstract	II
Acknowledgements	IV
List of Figures	IX
List of Tables	XII
1 Introduction	1
1.1 Motivation	1
1.2 Problem Background	2
1.3 Research Problems	4
1.4 Shortcomings of the Traditional Approach	5
1.5 Solutions	6
1.6 Summary of Contributions	7
1.7 Thesis Statement	10
1.8 Thesis Outline	10
2 Background	12
2.1 Reinforcement Learning	12
2.1.1 Q-learning	22
2.2 Reward Shaping	25
2.2.1 Psychological Perspective	26
2.2.2 Overview of Computational Shaping	26
2.2.3 Overview of Reward Shaping	29
2.2.4 Designing an Optimal Reward Function	31
2.2.5 Reward Shaping and Credit Assignment Problem	32

2.3	Categories of Human Feedback	34
2.3.1	Summary of Human Feedback Types: Used in this thesis . . .	34
2.3.2	Formal Categories	36
2.3.3	Preference-based Feedback	37
2.4	Related Studies: Learning from Human Feedback	39
3	Rewards from State Preferences	43
3.1	Introduction	43
3.2	Learning-to-Rank	45
3.2.1	Learning-to-Rank Approaches	47
3.2.2	Ranking SVM	50
3.3	Potential-Based Reward Shaping	53
3.3.1	Motivation	54
3.3.2	An Intuitive Solution	54
3.3.3	Formalization	55
3.3.4	Potential Function Design	59
3.3.5	Extensions and Related Work	62
3.4	Potential-Based Reward Shaping using Rewards from State Preferences	65
3.4.1	Formulation	67
3.4.2	Derivation of Mapping Function	70
3.4.3	The Proposed Algorithm	72
3.5	Experiments & Results	73
3.5.1	Comparison with a Baseline RL	73
3.5.2	State Preferences vs Reward Signals	75
3.6	Discussion	79
3.6.1	Overcoming Ranking SVM's Shortcomings	79
3.6.2	Access to Complete State Space	80
3.6.3	Tackling Equipotent Actions	80
3.6.4	Implementation Considerations	81
3.6.5	Visually Similar States	81
3.6.6	Use of Exploratory Policy	83
3.7	Extensions and Future Directions	83
3.7.1	From State Preferences to State-action Preferences	83
3.7.2	Three or More States in a Query	84
3.7.3	Other Research Directions	84
3.8	Conclusion	86
4	Rewards from Action Labels	88
4.1	Introduction	88
4.2	Use of Demonstrations	91
4.3	Learning from Demonstrations vs Interactive Reinforcement Learning	91

4.4	Need for Human-Generated Rewards	92
4.4.1	Assumptions on Human Trainers	93
4.5	Potential-Free Reward Shaping	94
4.5.1	Formal Definition	96
4.5.2	Related Studies	99
4.6	Risk of Positive Rewards Circuits	103
4.7	Learning Optimality and Termination	105
4.8	Rewards from Scalar Feedback	108
4.8.1	Formulation and Algorithm	110
4.9	Rewards from Action Labels	110
4.9.1	Formulation and Algorithm	114
4.10	Test Domain	116
4.10.1	Complex Scenarios	117
4.11	Experiment I	118
4.11.1	Experimental Setup	119
4.11.2	Results	120
4.11.3	Offline Performance Results	125
4.12	Experiment II	126
4.12.1	Setup	126
4.12.2	Results	129
4.13	Discussion on Rewards from Action Labels	135
4.13.1	Applications and Implementation Considerations	136
4.13.2	A Use Case of Predictive Action Model	138
4.13.3	Multiple Optimal Policies	138
4.13.4	Extensions and Further Studies	139
4.14	Conclusion	140
5	Rewards from Part-time Trainers	142
5.1	Introduction	142
5.2	Learning from Online Crowd-Workers	144
5.3	Learning Mechanism	145
5.4	Framework for rewards from Full-time Trainer	146
5.4.1	Definition of Full-time Trainer	147
5.5	Framework for rewards from Part-time trainers	147
5.5.1	Definition of Part-time Trainers	148
5.5.2	Framework	149
5.6	Experiments	149
5.6.1	Experimental Design	149
5.7	Results	153
5.7.1	Data of the Learned Policies	154
5.7.2	Online Performance	155

5.7.3	Offline Performance	157
5.7.4	Rewarding behavior	159
5.7.5	Survey Results	161
5.8	Discussion on Equivalence of the Learned Policies	162
5.9	Further Experiments	164
5.10	Conclusion	165
6	Conclusion	167
A	Support Vector Machine	173
B	Proof of Theorem 3.1	176
B.0.1	Proof of Sufficiency	176
C	Supplementary Material for Chapter 5	178
	Bibliography	181

List of Figures

1.1	The hierarchy of the topics presented in this document.	7
2.1	Sequential decision making process	13
2.2	A typical structure of a reinforcement learning paradigm (Sutton and Barto (1998), Figure 3.1).	14
2.3	An example of Q-values learned in a 3x4 grid-world domain. Each square represents a state. Each triangle contains a Q-value i.e. value for taking respective action from a state.	22
2.4	An overview of information flow in reward shaping. It shows a supervisor-in-the-loop scenario. The supervisor can be human or non-human. . .	30
3.1	Document retrieval process	46
3.2	A learning-to-rank framework	48
3.3	An example of ranking using weight vector in a 2-dimensional space. Two different weight vectors w_1 and w_2 produce two different rankings of four points.	51
3.4	8x8 grid-world domains with start state at the bottom-left corner and the goal state at the top-right corner. Left: A nice potential function defined over state-space using the Manhattan distance of a state from the goal state. Right: The same grid-world and potential function but with a wall at the center denoted by 'X'.	60
3.5	Maze domain	66
3.6	Perfect ranking	66
3.7	Overview of potential-based reward shaping using rewards from state preferences.	66
3.8	The process of extracting potential function from state preferences. .	66
3.9	An example of a query. The input records which state user prefers for the given goal state.	68
3.10	Average number of swapped pairs.	74
3.11	Average accumulated rewards per episode.	74
3.12	Average accumulated rewards per episode for reward shaping using feedback as reward signals.	76

3.13	Average accumulated rewards per episode for pbrs-rfsp using reward style feedback.	78
3.14	Average accumulated rewards per episode.	79
3.15	Images showing a possible set of state representation for the task of pick-and-place. The sub-captions indicate the possible preceding actions.	82
4.1	Two simple scenarios showing the optimal shaping rewards required for each action to ensure potential-based like reward shaping.	106
4.2	An example of information flow for reward shaping using human-supervisor's scalar feedback. Different types of scalar feedback are also shown.	109
4.3	An example of information flow for reward shaping using human-supervisor's scalar feedback. Different types of scalar feedback are also shown.	111
4.4	The testbed domain of Sokoban. From left to right it shows the result of two consecutive down actions.	116
4.5	Two difficult scenarios for non-expert users. The top row shows a wrong (sub-optimal) policy and the bottom row shows an optimal policy.	118
4.6	Snapshot of the desktop application designed to provide an interface for user experiments.	119
4.7	Comparison of SfID with SfIR. (a) Absolute temporal difference error as given by Equation 4.8. (b) The steps taken in each episode. (c) Cumulative discounted reward earned over the episodes. (d) The count of teachers input in each episode, (reward signal in interactive reward shaping and demonstration in SfID). The x-axis represents the number of episodes in each case. The labels of the y-axes are the respective headings of the sub-figures.	122
4.8	Offline performance results for policies learned via SfID and SfIR using feedback from a naïve user and an expert.	125
4.9	Instructions provided to the naïve users on AMT before they accepted to take part in the study.	127
4.10	Instructions provided to the users on AMT to complete practice for training using SfIR. It also shows the interface used to provide positive and negative rewards.	128
4.11	Instructions provided to the users on AMT to train the agent using SfIR. It also shows the interface used to provide positive and negative rewards.	129
4.12	Instructions provided to the users on AMT to practice training the agent using SfID. It also shows the interface used to provide the action labels.	130

4.13	Instructions provided to the users on AMT to actually train the agent using Sfid. It also shows the interface used to provide the action labels.	131
4.14	Results for the first question of the survey.	132
4.15	Results for the second question of the survey.	133
4.16	Results for the third question of the survey.	133
4.17	Results for the fourth question of the survey.	133
4.18	Results for the fifth question of the survey.	134
4.19	Results for the sixth question of the survey.	135
4.20	(a) State before action. (b) State after action.	138
5.1	A simple scenario of single trainer providing numeric rewards to each action.	147
5.2	A scenario of multiple trainers providing numeric rewards to each action for a fixed number of steps.	150
5.3	The proposed framework to sequentially learn a policy	151
5.4	Training phase web interface	152
5.5	Comparison of the learning performance of the crowd policies and the expert policy	156
5.6	Offline test performance for three crowd policies, individual teacher's policy and simulated expert policies	158
5.7	Positive versus negative rewards by part-time trainers for three policies learned via rewards from part-time trainers framework. The horizontal axis represents the number of episodes.	161
5.8	Comparison of positive versus negative rewards by part-time trainers (averaged for the three policies learned via rewards from part-time trainers framework) and by a full-time trainer. The horizontal axis represents the number of episodes.	161
5.9	Survey results: (a) To what length a person can comfortably teach this agent as a PAID task? (b) To what length a person can comfortably teach this agent as an UNPAID task?	162
A.1	An example of hyperplane separating points in two dimension (left). A depiction of geometric margin between points A and B (right). . .	174
A.2	Hard-margin SVM for separable data, with functional margins of 1 (left). Soft-margin SVM for unseparable data(right).	174
C.1	The survey questionnaire used to collect qualitative feedback from the AMT workers.	178
C.2	Snapshot of the HIT presented to the AMT workers. It shows the information provided to a workers before they accepted the HIT. . . .	179
C.3	A snapshot of the practice phase used to familiarize the user with the dynamics of the game.	180

List of Tables

1.1	Summary of the proposed method of generating rewards from human feedback. It compares the three methods for three properties.	9
2.1	Summary of the input features of the proposed method of generating rewards from human feedback.	35
2.2	Categorization of the feedback method for the three methods of learning from human feedback proposed in this work based on the categories proposed in literature.	37
3.1	Mean and standard deviation values for two measures of performance: Model error and total rewards	73
3.2	Mean and standard deviation values of total rewards for comparison of PBRS-RfSP with direct reward signals approach.	76
4.1	P-values for the absolute error learning curves shown in Figure 4.7a. We used t-Test for two samples assuming unequal variances. The values are reported for $P(T \leq t)$ one-tail/ $P(T \leq t)$ two-tail.	123
4.2	Total of the four learning performance measures summed over 30 episodes. We rounded-off absolute error and cumulative rewards values to improve readability.	124
4.3	The states and state-action pairs explored by the policies.	126
4.4	The average and variance values for the 21 policies trained by the naive users for both SfIR and SfID.	130
4.5	Examples of simple and composite actions along with some possible domains or applications. The ‘+’ sign means two actions are performed together and ‘→’ indicates a following action.	136
5.1	Comparison of the learned policies based on various important properties of a policy.	154
5.2	The percentages of correct input provided by the part-time and full-time trainers for the respective policies they taught. The correctness of an input was based on its accordance with an expert policy.	154

5.3	The percentages of correct input provided by the part-time and full-time trainers for the respective policies they taught. The correctness of an input was based on its accordance with an expert policy.	158
5.4	Positive and negative rewards provided during training of three policies learned via rewards from part-time trainers framework and one policy taught by a full-time trainer.	159

Dedicated to my family and friends